EXPRESS MAIL CERTIFICATE
Date 3/14/01 Label No. EL706722751US
I hereby certify that, on the date indicated above, this paper or fee was deposited with the U.S. Postal Service & that it was addressed for delivery to the Assistant Commissioner for Patents, Washington, DC 20231 by "Express Mail Post Office to Addressee" service.
D Breck
Name (Print) Signature

Nack Suppression for Multicast Protocols in Mostly One-Way Networks

BACKGROUND OF THE INVENTION

1. Field of the Invention.

This invention relates to multicast transmission of information across a data network. More particularly this invention relates to a technique for suppressing requests for retransmission of missing information by recipients at a downstream link of a mostly one-way data network.

2. Description of the Related Art.

The push model for distributing data over the Internet and other client server networks has become more widespread in recent years. In modern versions of this model a server "multicasts" data to an interested subset of clients on the network, known as a "multicast group". Whoever is interested becomes a listener by joining the group.

By their nature, push applications are closer to the broadcasting paradigm of radio and television than to the interactive paradigm of the World Wide Web. As such, broadband networks, such as cable TV or satellite, can be used as a very efficient medium for the transmission of "pushed data". Unfortunately, currently these networks are one-way only. That is to say, data such as a television program is sent from a broadcasting facility (the head-end) to several receivers (end-users) without any

feedback. As such, these networks are inappropriate for popular interactive push applications since the latter require a return channel. Although attempts to upgrade the current public network infrastructure are underway in several places around the world, it will take some years until reliable two-way broadband networks are commonplace and therefore, a mechanism for multicasting over one-way broadband networks is desirable.

In conventional unicast transmission, there are well known handshaking protocols that insure that blocks of data have been correctly received by the downstream recipient. The sender is easily able to determine whether a block of data requires retransmission, and can maintain awareness of the state of the recipient with little computational overhead. Without modifications, the unicasting techniques become unwieldy when applied to multicast transmission. Large numbers of recipients asynchronously requesting retransmissions are likely to produce congestion of the data network and degrade its performance. Furthermore, in the case of hierarchical multicasting arrangements, the algorithms based on estimating the paths to the recipient do not generalize well to tree structures, and the model of the communication as a conversation between sender and recipient does not stand up well.

Various techniques have been attempted to minimize requests for retransmission, such as the formation of token rings, and the use of a central controller. However

these techniques have had limited success due to problems of scalability, and the need to secure group-wide agreement. The latter approach is difficult in the case of groups having transient membership.

In the document, *A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing,* Sally Floyd *et al.,* SIGCOMM '95, Cambridge MA (1 Sept. 1995), it is proposed to employ a model known as Application Level Framing. This is a decentralized approach, in which all traffic is multicast. Each session has a bandwidth limit, below which group members send multicast reports advising of their current state. Receivers learn of missing data either by examination of their received sequences, or by evaluation of a multicast report from another group member. Repair requests are multicast among the group. When a node is in receipt of a request or reply concerning a data element which it lacks, it suppresses its own request for repair. After receiving content from the server 10, the receivers 12, 14, 16 may transmit acknowledgements or negative acknowledgements reflecting their current state. The reports are limited such that they do not exceed more than a predetermined percentage of the network's bandwidth, in order to prevent network congestion. Nevertheless, there is necessarily some degradation of network performance

The document, *Reliable Multicast Transport Protocol,* Shioshita, Teruji *et al.,* Draft Document for the 37th

IETF, Feb. 7, 1997, proposes a transport control mechanism to enable reliable multicast data transfer to a large number of receivers on a TCP/IP network from a server in parallel. This protocol promotes short delivery time, as the data is transferred only once, and conserves bandwidth because only one copy of the data is sent to the server. It has the advantage of requiring only a single session regardless of the number of receivers. However, despite some optimizations, there remains a requirement for receiver confirmation by ACK/NAK responses and the retransmission of data to selected receivers based on the information associated with the NAK response are disadvantages, as large numbers of receivers issuing ACK/NAK responses can still cause network congestion.

Another known multicast transport protocol is proposed in *Starburst Multicast File Transfer Protocol (MFTP) Specification.* Miller, K. *et al.,* Internet Draft, April 1998. This protocol operates in the Application Layer.

In copending U.S. Application No. 09/138,994, filed Aug. 24, 1998, of common assignee herewith, and hereby incorporated by reference, a technique of IP multicasting over existing broadband networks without using a return link is disclosed. This technique allows the issues of multicast group membership and error detection and recovery to be handled locally within an end-user terminal, without need for returning data to a host. According to

the technique a single data transmitter sends a group of data items to a subset of possible receivers over a one-way channel. Each data item is divided into blocks which are encapsulated to form datagrams, each including a block sequence number, a data item identifier, and a timestamp indicating the age of the data item. A group directory is regularly sent by the transmitter to each of the possible receivers. The group directory contains information for all groups of data items, enabling each receiver to select the group of data item it wishes to receive. Reliability is provided by periodic retransmission of missing data. Despite these advantages, significant problems remain.

SUMMARY OF THE INVENTION

It is a primary advantage of some aspects of the present invention that missing elements in a multicast transmission can be supplied while processing a minimum number of repair requests from clients.

It is another advantage of some aspects of the invention that network congestion is minimized during multicasting.

It is yet another advantage of some aspects of the invention that network traffic is optimized during multicast transmission.

These and other advantages of the present invention are attained by a multicasting system suitable for use in data network caches, software distribution arrangements,

and content provider applications in general. Content is multicast from a sender to a plurality of receivers over a data network, such as the Internet. Each receiver independently determines whether it is missing elements or packets of the content. Receivers having missing content each initiate a random timer. The receiver having the shortest random interval unicasts a negative acknowledgement to the sender. The sender immediately multicasts the negative acknowledgement to the other receivers. All other receivers having the same missing packet thereupon suppress their own negative acknowledgements as to that packet. A repair transmission is then multicast by the sender to all receivers. In some embodiments, the repair transmission could be multicast by a receiver possessing the missing packet. The random intervals have upper and lower bounds according to the round trip transmission time and the size of the largest missing data element.

The invention provides a method of transmitting data over a communications network, which includes multicasting content in a first multicast over a data network from a sender to a multicast group. The group comprises a plurality of receivers. Concurrently, in each of the receivers the method includes detecting a missing portion of the content, and delaying for a random interval. Thereafter no more than one negative acknowledgement is sent in a second transmission from one of the receivers to the sender. Thereafter, responsive to the negative acknow-

ledgement the missing portion is multicast in a third multicast from the sender to the multicast group.

According to an additional aspect of the invention, the random interval has a lower limit given by

$$LL = (a_1 t_{min}) x\ b$$

wherein x is a multiplication operator, a_1 is a proportionality constant, t_{min} is the minimal round trip transmission time between the sender and a respective one of the receivers, and b is a size of a largest packet of the missing portion.

According to an aspect of the invention, the random interval has an upper limit given by

$$UL = (a_1 t_{min}) x\ b$$

wherein x is a multiplication operator, a_2 is a proportionality constant, t_{max} is the maximum round trip transmission time between the sender and a respective one of the receivers, and b is a size of a largest packet of the missing portion.

Another aspect of the invention includes determining a current quantity of traffic on the data network, wherein the second multicast is sent when the current quantity is less than a predetermined value.

According to a further aspect of the invention, the random interval is the shortest random interval of the receivers.

According to yet another aspect of the invention, the third multicast is sent by the sender.

According to still another aspect of the invention, the third multicast is sent by one of the receivers.

The invention provides a computer software product, comprising a computer-readable medium in which computer program instructions are stored, which instructions, when read by a computer, cause the computer to execute a method of transmitting data over a data network. The method includes multicasting content in a first multicast over the data network from a sender to a multicast group comprising a plurality of receivers. Concurrently, in each of the receivers the method includes detecting a missing portion of the content, and delaying for a random interval. Thereafter no more than one negative acknowledgement is transmitted to the sender. Thereafter, responsive to the negative acknowledgement, the sender repeats the negative acknowledgement to the receivers, others of which thereupon suppress their own negative acknowledgements. The missing portion is then multicast in a third multicast from either the sender or another of the receivers to the multicast group.

The invention provides a computer system, which includes a first computer, and a second computer interconnected in a data network with the first computer. The first computer and the second computer receive multicast content in a first multicast via the data network from a content server. The first computer and the second computer have program instructions stored therein, which in-

structions cause the first computer and the second computer to concurrently execute a method of transmitting data over the data network. The method includes detecting a missing portion of the content, determining random intervals, wherein the random interval of the first computer is shorter than the random interval of the second computer. The first and second computer delay for their respective random intervals. Thereafter the first computer sends a first negative acknowledgement to the content server. In the event that the second computer has not received the missing portion, the second computer suppresses a second negative acknowledgement therefor. The missing portion is contained in a third multicast from the content server.

According to an aspect of the invention, the third multicast is sent by the second computer.

BRIEF DESCRIPTION OF THE DRAWINGS

For a better understanding of these and other objects of the present invention, reference is made to the detailed description of the invention, by way of example, which is to be read in conjunction with the following drawings, wherein:

Fig. 1 is a schematic illustrating the interconnection of a computer system in a data network for use according to the present invention; and

Fig. 2 is a flow chart of a method of transmitting data according to the invention.

DESCRIPTION OF THE PREFERRED EMBODIMENT

In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent however, to one skilled in the art that the present invention may be practiced without these specific details. In other instances well known circuits, control logic, and the details of computer program instructions for conventional algorithms and processes have not been shown in detail in order not to unnecessarily obscure the present invention.

Software programming code, which embodies the present invention, is typically stored in permanent storage of some type, such as a computer readable medium. In a client/server environment, such software programming code may be stored on the client or a server. The software programming code may be embodied on any of a variety of known media for use with a data processing system, such as a diskette, or hard drive, or CD-ROM. The code may be distributed on such media, or may be distributed to users from the memory or storage of one computer system over a network of some type to other computer systems for use by users of such other systems. The techniques and methods for embodying software program code on physical media and/or distributing software code via networks are well known and will not be further discussed herein.

Turning now to the drawings, and to Fig. 1 thereof, there is shown a high level architectural diagram of a preferred embodiment of a multicasting system 18 which employs the techniques of the present invention. The source of the content to be distributed is a server 10, which is provided with transmitting capability, and also has receiving capability. In a hierarchical tree, the server 10 is a parent with respect to a plurality of downstream receivers 12, 14, 16. Content is multicast from the server 10 via a data network, which may be the Internet, to the receivers 12, 14, 16. The receivers 12, 14, 16 are generally provided with transmitting capability also, and are interconnected with one another and with the server 10. The receivers 12, 14, 16 are linked respectively to random timing circuits 20, 22, 24, which trigger transmissions by the receivers 12, 14, 16 in a manner disclosed hereinbelow.

The server 10 preferably employs the REMADE protocol to multicast the content. The REMADE protocol is disclosed in the above noted U.S. Application No. 09/138,994. The REMADE protocol is a technique of IP multicasting over existing broadband networks without using a return link. This technique allows the issues of multicast group membership and error detection and recovery to be handled locally within an end-user terminal, without need for returning data to a host. According to the technique, a single data transmitter sends a group of data

items to a subset of possible receivers. Each data item is divided into blocks, which are encapsulated to form datagrams, each including a block sequence number, a data item identifier, and a timestamp indicating the age of the data item. A catalog, comprising a group directory is regularly sent by the transmitter to each of the possible receivers. The group directory contains information for all groups of data items, enabling each receiver to select the group of data item it wishes to receive. Reliability may be enhanced by periodic retransmission of missing data.

In some embodiments, improvements in the REMADE protocol which were disclosed in our copending Application No. 09/564,387, filed May 3, 2000, and herein incorporated by reference, may be used in the practice of the present invention.

The procedure for dealing with negative acknowledgements is now explained with reference to Fig. 1 and Fig. 2. In the configuration of the data network, the minimal and maximal round trip times between the server 10 and each of the receivers 12, 14, 16 is determined by known techniques. The multicast transmission occurs in blocks or packets, which may vary in size, according to the particular transmission protocol employed. However, the size of the blocks or packets is also known to the receivers 12, 14, 16.

At initial step 26 the server 10 multicasts content, for example a group directory, in accordance with the REMADE protocol. After receiving content from the server 10 at step 28, the receivers 12, 14, 16 then evaluate the content to determine whether data is missing at decision step 30.

If at decision step 30 it is determined that no data is missing, then no action is required, and control proceeds to termination step 32.

If at decision step 30 it is determined that data is missing, for example in the receiver 12, then at step 34 the random timing circuit 20 is initiated. A delay occurs for a random interval, until a signal is received by the receiver 12 from the random timing circuit 20. The random interval has a lower limit given by

$$LL = (a_1 t_{min}) \times b \qquad (1)$$

wherein "x" is the multiplication operator. The factor a_1 is a proportionality constant. The value t_{min} is the minimal round trip transmission time between the server 10 and the receiver 12. The value a_1 may be adjusted for a particular application, or in some embodiments can be varied according to prevailing conditions on the data network. The quantity b is the size of the largest missing data element.

The random interval has an upper bound given by

$$UL = (a_2 t_{max}) \times b \qquad (2)$$

wherein "x" is the multiplication operator. The factor a_2 is a proportionality constant. The value t_{max} is the maximum round trip transmission time between the server 10 and the receiver 12. The value a_2 may be adjusted for a particular application, or in some embodiments can be varied according to prevailing conditions on the data network. The quantity b is the size of the largest missing data element.

When a signal is received from the random timing circuit 20, the receiver 12 transmits a negative acknowledgement reflecting its current state at step 36 to the sender. In some embodiments, the transmission of step 36 could be multicast to the other receivers as well. This transmission is inhibited, however, if it would cause network traffic to exceed more than a predetermined percentage of the network's bandwidth. This limitation is desirable in order to minimize network congestion. It is understood that decision steps 30 and step 34 are also occurring simultaneously in the receivers 14, 16, but that the random timing circuit 20 was the first of the random timing circuits 20, 22, 24 to trigger.

When the receiver 12 transmits its current state, the report is received by the server 10 at step 38. At step 40, the server repeats the negative acknowledgement to the other receivers in a multicast report. At step 42 the report is received by the receivers 14, 16, as indicated by the arrows in Fig. 1.

At decision step 44 the receivers 14, 16 independently evaluate whether they are missing the same information as was reported missing by the receiver 12 in step 36. If the result of this test is affirmative, then, at step 46, any of the receivers 14, 16 missing the information simply wait until a repeat transmission is sent by the server 10. As shown at step 45, they do not multicast or otherwise transmit their own negative acknowledgement as to the information which was reported as missing by the receiver 12.

It the result of the test at decision step 44 is negative, then the receivers 14, 16 simply resume awaiting the random timing circuits 22, 24 to trigger, and control proceeds to termination step 32.

In response to reception of the multicast of the receiver 12 at step 38, and following completion of step 40, a repair transmission is multicast at step 48. The repair transmission is effected in some embodiments by the server 10. In other embodiments one of the receivers 14, 16 may respond, according to a particular control policy. A control policy, for example, could require the first unit receiving the negative acknowledgement from the receiver 12 to produce a repair transmission if possible.

The repair transmission is detected by the receivers 12, 14, 16. All further repair transmissions are then suppressed. Those receivers, which are missing the data

element contained in the repair transmission, accept it, and make appropriate internal corrections to their files at step 50. Control then proceeds to termination step 32.

In the foregoing description it could be the case that all three of the receivers 12, 14, 16 were missing a particular data element. Yet, only one negative acknowledgement was placed on the data network. Up to two messages were suppressed, with corresponding benefit to the load on the data network.

It should be noted that while the system 18 is represented for clarity in Fig. 1 as a two-level tree, comprising the server 10 and the receivers 12, 14, 16, there can be any number of levels arranged in a more complex tree-structured hierarchy, as appropriate for a particular application. In such case the receivers 12, 14, 16 communicate with a lower level of receivers, which may be part of the same or a different multicast group. The improvement in performance of the data network linking all the receivers is scalable, and the advantage of the disclosed arrangement as compared with the conventional multicasting systems becomes greater as the number of receivers increases.

While this invention has been explained with reference to the structure disclosed herein, it is not confined to the details set forth, and this application is intended to cover any modifications and changes as may come within the scope of the following claims: